TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI:**

**DỰ ĐOÁN HẠT GIỐNG DỰA THEO 7 THUỘC TÍNH**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện:

1. Nguyễn Quang Tú - 1951061091

2. Đoàn Hùng Dương - 1951060650

3. Trần Văn Tuấn - 1951010147

**Hà Nội, năm 2023**

**MỤC LỤC**

**Phần 1: Tổng quan**

***1. Giới thiệu về học máy***

* Lịch sử và vai trò của machine learning
* Lịch sử của Machine Learning:
* 1950 – Alan Turing tạo ra phép thử Turing (Turing Test) để xác định xem một máy tính có trí thông minh thực sự hay không. Để vượt qua các phép thử, máy tính phải có khả năng đánh lừa một người tin rằng nó cũng là con người.
* 1952 – Arthur Samuel đã viết chương trình máy học đầu tiên. Chương trình này là trò cờ đam, và máy tính IBM càng chơi càng tiến bộ, học được các nước đi tạo lập nên chiến lược để chiến thắng và tích hợp các nước đi đó vào chương trình của mình.
* 1957 – Frank Rosenblatt thiết kế hệ thần kinh (neural network) đầu tiên cho máy tính (gọi là “perceptron”) mô phỏng các quá trình tư duy của não người.
* 1967 – Thuật toán “điểm lân cận gần nhất” (nearest neighbor) được viết ra cho phép các máy tính bắt đầu sử dụng dạng mẫu nhận thức rất cơ bản. Thuật toán này có thể được sử dụng để vẽ một tuyến đường cho nhân viên bán hàng lưu động, bắt đầu từ một thành phố ngẫu nhiên nhưng đảm bảo họ tới tất cả các thành phố trong một chuyến đi ngắn.
* 1979 – Sinh viên Đại học Standford phát minh ra xe “Standford Cart” có thể tự xác định được vật cản trong phòng.
* 1981 – Gerald Dejong giới thiệu khái niệm Học qua giải thích (Explanation Based Learning – EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung mà nó có thể làm theo bằng cách loại bỏ các dữ liệu không quan trọng.
* 1985 – Terry Sejnowski phát minh ra NetTalk, học cách phát âm các giống như một đứa trẻ.
* 1990s – Các nghiên cứu về marchine learning chuyển từ cách tiếp cận hướng-kiến-​​thức sang cách tiếp cận hướng-dữ-liệu. Các nhà khoa học bắt đầu tạo các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra kết luận – hay “học” – từ các kết quả.
* 1997 – Máy tính Deep Blue của IBM đánh bại nhà vô địch cờ vua thế giới.
* 2006 – Geoffrey Hinton tạo ra thuật ngữ “học sâu” (deep learning) để giải thích các thuật toán mới cho phép các máy tính “nhìn thấy” và phân biệt các đối tượng và văn bản trong hình ảnh và video.
* 2010 – Máy Microsoft Kinect có thể theo dõi 20 tính năng của con người ở tốc độ 30 lần mỗi giây, cho phép mọi người tương tác với máy tính thông qua các động tác và cử chỉ.
* 2011 – Máy IBM Watson đánh bại các đối thủ con người tại Jeopardy.
* 2011 – Google Brain được phát triển, và hệ thần kinh sâu (deep neural network) của nó có thể học để phát hiện và phân loại các đối tượng bằng với khả năng của một chú mèo.
* 2012 – X Lab của Google phát triển một thuật toán machine learning có khả năng tự duyệt video YouTube để xác định các video có chứa hình ảnh mèo.
* 2014 – Facebook phát triển DeepFace, một thuật toán phần mềm có thể nhận diện hoặc xác minh các cá nhân trong các bức ảnh ở mức độ ngang với con người.
* 2015 – Amazon ra mắt nền tảng machine learning riêng của mình.
* 2015 – Microsoft tạo ra bộ công cụ máy học được phân bổ (Distributed Machine Learning Toolkit), cho phép phân bổ hiệu quả machine learning trên nhiều máy tính
* 2015 – Hơn 3.000 nhà nghiên cứu về AI và Robotics, được Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác) hậu thuẫn, đã ký một bức thư ngỏ cảnh báo nguy cơ của vũ khí tự hành có thể lựa chọn và nhắm vào các mục tiêu mà không cần sự can thiệp của con người.
* 2016 – Thuật toán trí tuệ nhân tạo (artificial intelligence algorithm) của Google đã đánh bại một cầu thủ chuyên nghiệp trong trò cờ Go của Trung Quốc, vốn được coi là trò cờ phức tạp nhất thế giới và khó hơn cờ vua gấp nhiều lần. Thuật toán AlphaGo do Google DeepMind phát triển đã giành chiến thắng năm trên năm trận đấu Go.
* Vai trò của Machine Learning:
* Tập trung phát triển các chương trình máy tính có khả năng truy cập DL để sử dụng trong việc tự học
* Machine Learning có thể xử lý hình ảnh bằng cách phân tích thông tin từ hình ảnh hoặc thực hiện 1 số phép biến đổi
* Khai thác dữ liệu
* Phân tích văn bản
* Ưu điểm và hạn chế của Học có giám sát và không giám sát
* Học máy có giám sát:

|  |  |
| --- | --- |
| Ưu điểm | Hạn chế |
| * Cho phép thu thập dữ liệu và tạo ra dữ liệu đầu ra từ những kinh nghiệm trước đó * Tối ưu hóa tiêu chí hiệu suất với sự trợ giúp của kinh nghiệm * Phương pháp: Đơn giản * Độ chính xác: Cao | * Phân loại dữ liệu lớn là 1 thách thức * Độ phức tạp tính toán: Cao |

* Học máy không giám sát:

|  |  |
| --- | --- |
| Ưu điểm | Hạn chế |
| * Sử dụng bộ dữ liệu không có kết quả được gán nhãn để dự đoán kết quả của dữ liệu không nhìn thấy * Độ phức tạp tính toán: Thấp | * Phương pháp: Tính toán phức tạp * Độ chính xác: Thấp |

***2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn***

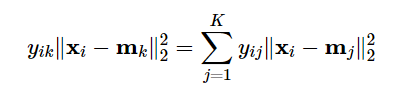
a. Phương pháp clustering

* Mục đích của phương pháp: phân các điểm gần nhau thành 1 cụm
* Input: Dữ liệu chưa được gán nhãn, và số lượng cụm muốn tìm kiếm
* Output: Chỉ ra center mỗi nhóm và phân các điểm dữ liệu vào các nhóm
* Cách thực hiện:

Nếu ta coi center mk là center (hoặc representative) của mỗi cluster và ước lượng tất cả các điểm được phân vào cluster này bởi mk, thì một điểm dữ liệu xi được phân vào cluster k sẽ bị sai số là (xi−mk)(xi−mk). Chúng ta mong muốn sai số này có trị tuyệt đối nhỏ nhất nên ([giống như trong bài Linear Regression](https://machinelearningcoban.com/2016/12/28/linearregression/#sai-so-du-doan)) ta sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:



Hơn nữa, vì xi được phân vào cluster k nên yik=1, yij=0, yik=1, yij=0, ∀j≠k. Khi đó, biểu thức bên trên sẽ được viết lại là:



Sai số cho toàn bộ dữ liệu sẽ là:

A picture containing text, watch

Description automatically generated

Trong đó Y=[y1;y2;…;yN]Y=[y1;y2;…;yN], M=[m1,m2,…mK]M=[m1,m2,…mK] lần lượt là các ma trận được tạo bởi label vector của mỗi điểm dữ liệu và center của mỗi cluster. Hàm số mất mát trong bài toán K-means clustering của chúng ta là hàm L (Y, M) với ràng buộc như được nêu trong phương trình (1)(1).

Tóm lại, chúng ta cần tối ưu bài toán sau:

Text

Description automatically generated

* Đánh giá phương pháp:
  + Có 1 vài hạn chế của thuật toán k-means clustering
    - Cần biết số lượng cluster cần clustering, không phải lúc nào cũng xác dịnh đc giá trị này
    - Nghiệm cuối cùng phụ thuộc vào các centers được khởi tạo ban đâu, Tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm
    - Các cluster có số lượng điểm gần bằng nhau
    - Các cluster cần có dạng hình tròn
    - Khi một cluster nằm phía trong 1 cluster khác thì k-mean cluster không thực hiện được

***3.Trình bày bài toán***

* Mô tả bài toán: Phân loại hạt giống theo dữ liệu đã cho sẵn
* Input: Tập dữ liệu không nhãn gồm A, P, C, LK, WK, A\_Coef và LKG
* Ouput: Loại hạt giống (kama, rosa, canadian)

**Phần 2: Thực nghiệm**

***1.Mô tả tập dữ liệu***

* Mô tả tập dữ liệu của bài toán, dữ liệu gồm những chiều thông tin A, P, C, LK, WK, A\_Coef và LKG, có 211mẫu dữ liệu.
* Mô tả ma trận dữ liệu (X) gồm A, P, C, LK, WK, A\_Coef và LKG
* A : Diện tích

P : Chu vi

C : Độ cứng

LK : Chiều dài của hạt

WK : Chiều rộng của hạt

A\_Coef : Độ biến dạng

LKG : Độ dài rãnh nhân

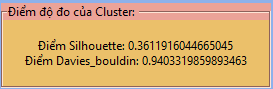
***2. Mô tả cách giải bài toán bằng phương pháp học máy***

Bước 1 :Dựa vào 2 độ đo **silhouette\_score** và **davies\_bouldin\_score** để xác định số cụm tốt nhất

Bước 2 :Dùng phương pháp Clutering để xây dựng mô hình và đánh giá mô hình .

***3. Đánh giá mô hình***

Dùng 2 độ đo silhouette\_score, davies\_bouldin\_score để đánh giá sự phù hợp của mô hình



***4. Mô tả các chức năng của chương trình***

-Chương trình bao gồm các labels mô tả thông tin, các ô nhập dữ liệu, group box thể hiện các độ đo và các button bấm để thể hiện dự đoán

**Kết luận**

- Tóm lược các nội dung chính mà bài tập lớn làm được

* Về mặt lý thuyết, bài tập lớn trình bày 2 phương pháp là K-means clustering
* Về mặt thực nghiệm, thuật toán Clustering. Sau đó áp dụng vào bài toán “Phân loại hạt giống” từ đó so sánh độ đo của 2 thuật toán và tìm thuật toán phù hợp với bài toán.

**Tài liệu tham khảo**

-Các tài liệu được tham khảo trong báo cáo

<https://tuanvanle.wordpress.com/2013/12/25/phuong-phap-phan-tich-thanh-phan-chinh-principal-component-analysis-pca/>